# SMall Molecule Growth 2001 (SMoG2001): An Improved Knowledge-Based Scoring Function for Protein−Ligand Interactions

Alexey V. Ishchenko[†] and Eugene I. Shakhnovich*

*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138*

*Received December 26, 2001*

Computational lead design procedures require fast and accurate scoring functions to rank millions of generated virtual ligands for protein targets. In this article, we present an improved version of the SMoG scoring function, called SMoG2001. This function is based on a knowledge-based approach—that is, the free energy parameters are derived from the observed frequencies of atom−atom contacts in the database of three-dimensional structures of protein−ligand complexes via a procedure based on statistical mechanics. We obtained the statistics from the set of 725 complexes. SMoG2001 reproduces the experimental binding constants of the majority of 119 complexes of the testing set with good accuracy. On similar testing sets, SMoG2001 performs better than two other widely used scoring functions, PMF and SCORE1(LUDI), and comparably to DrugScore. SMoG2001 poorly predicts the affinities of ligands interacting via quantum mechanical forces with metal ions and ligands that are large and flexible. We attribute significant improvement in accuracy over previous versions of the SMoG scoring function to a better description of the reference state—that is, the state of no interactions.

## Introduction

Prediction of binding affinities of protein−ligand complexes is the most challenging part of computational ligand design.[1,2] Free energy of ligand binding cannot at present be calculated exactly from statistical mechanics or quantum chemistry and thus is approximated in various scoring functions (also referred to as force fields or potentials). These force fields that are applied for scoring putative protein−ligand complexes differ in speed and accuracy depending on the nature and the degree of approximations involved. In drug design projects, computational methods[3] are generally used in two consecutive steps: lead generation and lead optimization. In lead generation, fast and sufficiently accurate scoring functions are necessary to identify several hits from among many chemical structures and conformations docked to a protein or generated by a de novo ligand design program. Lead optimization requires force fields in which accuracy is more important than speed since only few plausible ligands are scrutinized.

Force fields used in computational ligand design can be classified into three major categories: all atom force fields, empirical scoring functions, and knowledge-based scoring functions. All atom force fields (such as CHARMM,[4] AMBER,[5] OPLS,[6] MMFF,[7] MM3,[8] and others) are usually employed in free energy perturbation (FEP) methods,[9,10] which use molecular dynamics simulations to calculate differences between binding free energies of closely related ligands to one protein. These methods are in most cases the most accurate ones with which to compute binding free energies. They are, however, computationally costly and are limited to structurally similar ligands, differing usually in only one functional group. Therefore, FEP simulations are mostly used in the lead optimization.

In the empirical approaches (such as SCORE1(LUDI),[11] SCORE2(LUDI),[12] VALIDATE,[13] and ProteusScore[14]), binding free energy is approximated in a "master equation" as a sum of several functions corresponding to arbitrary enthalpic and entropic contributions. These functions use free parameters, which are optimized by maximizing the correlation between computed and experimental binding free energies of a set of protein−ligand complexes. Scoring functions based on master equations are fast and often employed in docking programs for lead generation.

One of the recently developed approaches is rooted in the so-called knowledge-based methods[15] (such as SMoG,[16,17] PMF,[18] BLEEP,[19] DrugScore,[20,21] and others[22,23]). These methods were first employed for protein folding studies and recently applied for computing binding free energies of ligands. Knowledge-based potentials (KBP) are derived from the structures of protein−ligand complexes using statistical mechanics. Binding free energy is represented as a sum of free energies (or, equivalently, potentials of mean force, also referred to as parameters of the potential) of interatomic contacts that are calculated from their frequencies in the database of the structures via statistical mechanical procedures. Force fields based on KBP are fast, comparable in speed to empirical scoring functions, and can be used for lead generation.

In principle, accuracy of knowledge-based scoring functions is at least comparable to that of empirical force fields. Empirical force fields are parametrized on small (~100) sets of complexes featuring often similar ligands and proteins that must have both structure and binding constant known. Therefore, these scoring functions can be biased toward specific structural motifs and their

* To whom correspondence should be addressed. Tel.: (617)495-4130. Fax: (617)384-9228. E-mail: eugene@belok.harvard.edu.

[†] Present address: Concurrent Pharmaceuticals, 1 Broadway 14th floor, Cambridge, MA 02142.

10.1021/jm0105833 CCC: $22.00
Published on Web 05/14/2002

transferability to protein−ligand complexes with different interaction patterns requires more study. In contrast, only structural information is necessary to derive knowledge-based parameters; there are more (∼1000) structurally diverse complexes available. Therefore, knowledge-based scoring functions can be less biased to certain types of protein−ligand complexes.

We have recently developed a de novo ligand design program, called CombiSMoG,[16,17,24] which uses a coarse-grained knowledge-based scoring function and a combinatorial small molecule growth algorithm. In our previous work,[25] we tested CombiSMoG's scoring function using a self-consistent procedure for deriving the potential from a set of fictitious "toy" ligands. We found that the accuracy of the potential depends crucially on the proper definition of the so-called reference state—that is, the state of no interactions.

In that work, we formulated an arbitrary "true potential" and used it along with the CombiSMoG's growth algorithm to construct ∼10 000 toy ligands for 14 structurally diverse proteins. We separated the "toy database" into two parts: (i) "training set", the set of the complexes used for derivation of the potential to be tested, and (ii) "testing set", the set of the complexes used for testing this potential. We then formulated several knowledge-based functions with the same interaction model as in true potential. We suggested the following criteria of successful performance of a scoring function: (i) its derived parameters should be equal or close to those of the true potential, and (ii) the difference between derived and true parameters should decrease with increasing size of the training set. In other words, the correlation coefficient between true and derived interaction scores of the toy complexes of the testing set should be high (as close to 1.0 as possible) and should increase with increasing size of the training set.

The original scoring function (SMoG96[17]) did not meet these criteria. The parameters of the potential extracted by SMoG96 were different than the true parameters, resulting in a low (∼0.2) correlation coefficient between the derived and the true scores. In contrast, the new function (we called it SMoG2001) gave a significantly higher correlation coefficient of 0.8. We proposed that a new, statistical mechanically correct definition of the reference state was responsible for improvement in the performance of this function for toy complexes.

Here, we apply the SMoG2001 scoring function to real protein−ligand complexes. We assumed a priori that the suggested reasons for the improvement of SMoG2001 over SMoG96 for the toy database would also be applicable to real complexes. We expected, however, a smaller degree of improvement, since for real structures the true interaction model is unknown and additional approximations are necessary.

We organize the paper in the following way: in the Materials and Methods section, we present and explain the equations by which the numbers of contacts in the database are related to free energies in SMoG96 and SMoG2001 scoring functions. Using these equations, we derive the parameters of the potential from the training set of 725 protein−ligand complexes contained in the Protein Data Bank (PDB)[26] and test the scoring functions on a structurally diverse set of 119 complexes whose experimental binding constants are known.

In the Results and Discussion section, we first show that SMoG2001 computes higher correlation coefficients and gives lower standard deviations between predicted and experimental binding affinities in the testing set and in subsets of protein−ligand complexes than does SMoG96. We then show that the definition of the reference state suggested in the previous work on toy complexes and incorporated in SMoG2001 is the main factor responsible for the observed improvement. Next, we address two questions of the quality of the SMoG2001 scoring function: (i) whether the training database is large enough for the derivation of statistically robust potential and (ii) whether the use of a statistical mechanical procedure to convert database statistics to free energies of contacts of different atom types is generally meaningful, i.e., whether it derives more accurate potential than a nonspecific pairwise contact function (free energies of all contacts are the same). We compare the performance of the SMoG2001 function with that of the other widely used force fields: PMF, DrugScore, and SCORE1 (LUDI). Finally, we discuss two subsets of the protein−ligand complexes for which SMoG2001 fails to predict the binding affinities correctly: (i) for the complexes of metalloproteases, SMoG2001 fails to account for quantum mechanical effects involved in metal−ligand coordination, and (ii) for the large flexible peptidomimetic ligands of endothiapepsin complexes, it does not account for ligand conformational entropy. We show that the incorporation of additional empirical terms counterbalancing the loss of entropy due to freezing of flexible bonds of ligands upon binding can improve the prediction for large peptidomimetic molecules. We conclude with the suggestions for the further development of KBPs.

## Materials and Methods

We use the pairwise approximation for the binding free energy of the protein−ligand complex; i.e., it is the sum over all atom−atom contacts between protein and ligand atoms

$$F = \sum_{p} \sum_{l} F(\sigma_{\mathrm{p}},\sigma_{\mathrm{l}})\Delta(\mathrm{p,l}) \qquad (1)$$

In this equation, $p$ denotes a protein atom of type $\sigma_p$, $l$ denotes a ligand atom of type $\sigma_l$, $\Delta(p,l)$ is the characteristic function of the contact (1 if atoms $p$ and $l$ are in contact and 0 otherwise), and $F(\sigma_p,\sigma_l)$ is the value corresponding to the potential of mean force of interaction between two atoms of given atom types. All protein and ligand atoms are assigned atom types that depend on their chemical properties (element, hybridization, polarity, hydrogen bond donor/acceptor, and charge).

**SMoG96.** In the SMoG96 function, DeWitte and Shakhnovich defined that two atoms are in contact if the separation between them is less than 5 Å. This distance is approximately equal to the sum of radii of first coordination shells of water (solvent) around two interacting atoms. When the contact is formed, the coordination shells are destroyed, and the water molecules are transferred into the free solvent, resulting in the net gain of solvent entropy; therefore, contribution of the solvent entropy is implicitly taken into account in this approach and $F(\sigma_p,\sigma_l)$ are the potentials of mean force.

The $F(\sigma_p,\sigma_l)$ values were calculated as

$$F(\sigma_p,\sigma_l) = -\ln\left[\frac{p(\sigma_p,\sigma_l)}{p^{\mathrm{ref}}}\right] \qquad (2)$$

where $p(\sigma_p,\sigma_l)$ denotes the measure of frequency of the contacts between atom types $\sigma_p$ and $\sigma_l$ in the training database and $p^{\text{ref}}$ is the probability of those contacts in the hypothetical reference state. The $p(\sigma_p,\sigma_l)$ values were calculated as

$$p(\sigma_p,\sigma_l) = \frac{N(\sigma_p,\sigma_l)}{N(\sigma_p)^{1/2}\,N(\sigma_l)^{1/2}} \tag{3}$$

where $N(\sigma_p,\sigma_l)$ is the number of $(\sigma_p,\sigma_l)$ contacts calculated from the database, $N(\sigma_l)$ is the number of ligand atoms of type $\sigma_l$ that make at least one contact with any protein atom, and $N(\sigma_p)$ is the number of protein atoms of type $\sigma_p$ that make at least one contact with any ligand atom. The reference state was defined as a complex of randomly connected protein atoms and randomly connected ligand atoms that do not interact, i.e., $F(\sigma_p,\sigma_l) = 0$ for all $(\sigma_p,\sigma_l)$ pairs. The simple approximation $p^{\text{ref}} = \langle p(\sigma_p,\sigma_l)\rangle_{(\sigma p,\sigma l)}$ was chosen for the probabilities of the reference state. In this equation as well as in the subsequent ones, we denote $\langle X(\sigma_p,\sigma_l)\rangle_{(\sigma p,\sigma l)}$ as an average of $X$ over all atom types.

**SMoG2001.** In the newer version of the potential,[25] SMoG2001, we redefined the reference state that ensures proper normalization of contact probabilities (sum of all values over atom types is equal to 1) and introduced two distance intervals ("bins") over which the contact statistics are computed. This resulted in the following formulation of the scoring function (eqs 4−7):

$$F = \sum_r \sum_p \sum_l F(r,\sigma_p,\sigma_l)\,\Delta(r,p,l) \tag{4}$$

where for each distance interval, $r$ is the outer radius of the bin and two atoms are defined to be in contact if the distance between them is in the bin $(r,r\text{-}\Delta r)$.

The detailed description of the hypothesis behind the SMoG2001 function and of the derivations of the equations is given in our previous work.[25] In summary, the values $F(r,\sigma_p,\sigma_l)$ of potential of mean force are calculated as a logarithm of the ratio of probabilities of contact formation in the complex and in the reference state:

$$F(r,\sigma_p,\sigma_l) = -\ln\left[\frac{N(r,\sigma_p,\sigma_l)}{C(r) \times \Omega(r,\sigma_p,\sigma_l) \times \sum_{\sigma_p}\sum_{\sigma_l} N(r,\sigma_p,\sigma_l)}\right] \tag{5}$$

where $N(r,\sigma_p,\sigma_l)$ is the number of contacts between atoms of given atom types computed from the database, and $C(r) \times \Omega(r,\sigma_p,\sigma_l)$ is the probability of the contacts in the reference state for a given distance interval. $C(r)$ is a normalization constant, and $\Omega(r,\sigma_p,\sigma_l)$ is the term accounting for composition of the database, i.e., number of atoms of the $\sigma_l$ and $\sigma_p$ types participating in the protein−ligand interaction. To derive the expression for $C(r)$, the average of the eq 5 is taken over all atom types resulting in eq 6:

$$\ln C(r) = \left\langle \ln \frac{N(r,\sigma_p,\sigma_l)}{\Omega(r,\sigma_p,\sigma_l) \times \sum_{\sigma_p}\sum_{\sigma_l} N(r,\sigma_p,\sigma_l)} \right\rangle_{(\sigma_p,\sigma_l)} + F_0(r) \tag{6}$$

where $F_0(r) = \langle F(r,\sigma_p,\sigma_l)\rangle_{(\sigma_p,\sigma_l)}$ is a free parameter equal to an average free energy for a given distance interval. $\Omega(r,\sigma_p,\sigma_l)$ is defined in the first approximation as

$$\Omega(r,\sigma_p,\sigma_l) = N(r,\sigma_p)^{\alpha}\,N(r,\sigma_l)^{\beta} \tag{7}$$

where $N(r,\sigma_p)$ is the number of protein atoms of type $\sigma_p$ in the interval $(r,r\text{-}\Delta r)$ that make at least one contact with any ligand atom, $N(r,\sigma_l)$ is the number of ligand atoms of type $\sigma_l$ in the interval $(r,r\text{-}\Delta r)$ that make at least one contact with any protein atom, and $\alpha$ and $\beta$ are free parameters.

We derive the scoring function as following (see Figure 1 for the schematic flowchart of the calculation). We first compute the $N(r,\sigma_p,\sigma_l)$, $N(r,\sigma_p)$, and $N(r,\sigma_l)$ from the PDB of protein−ligand complexes that constitute the training set. We then choose first approximation values for the four free parameters in our model: $\alpha$, $\beta$, $F_0(r)$, and $r$ (the outer radii of distance intervals). Using these quantities and eqs 5−7, we derive the $F(r,\sigma_p,\sigma_l)$ values. We apply the scoring function to calculate the binding scores of the complexes for which the experimental binding constants are known. These complexes form the testing set. We use the correlation coefficient between computed and experimental binding free energies of the testing set complexes as the measure of the performance of the scoring function.

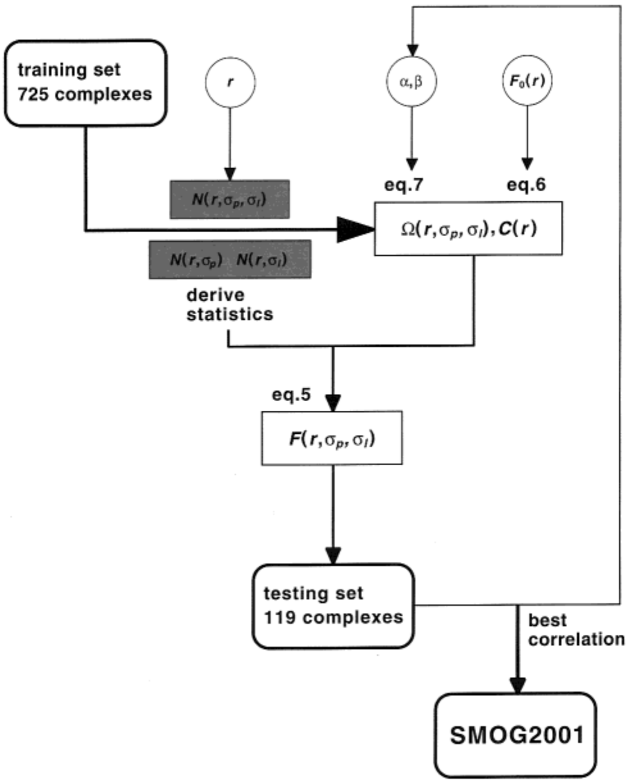

**Figure 1.** Flowchart showing the sequence of the derivation of the parameters of the potential from the database statistics. Thick arrows indicate the steps in which various quantities are computed corresponding to a given equation number; thin arrows show the relations between various quantities. Colored rectangles represent quantities derived from the database; rectangles, computed quantities; circles, free parameters.

We also investigate how the free parameters affect the correlation coefficient. In each of these calculations, we modify one parameter at a time within limited physical range and recompute the correlation coefficient. In all runs, we assumed $\alpha = \beta$; the accuracy of the scoring function does not depend on $F_0(r)$. Therefore, only two parameters are effectively adjusted in our scoring function. In the final version of the potential to be used for the design applications, we chose the values of the free parameters that gave the highest correlation coefficient.

The testing set consists of 119 complexes of eight structurally diverse protein subsets aspartic proteases (18 complexes), serine proteases (20 complexes), metalloproteases (22), carbonic anhydrase (19), sugar-binding proteins (14), endothiapepsin (11), purine nucleoside phosphorylase (5), and other proteins (10). The list of complexes and their binding constants are given in Table 1.

We calculate the number of contacts from the training set of 725 protein−ligand complexes deposited in the PDB. This set does not contain complexes with metal ions, inorganic fragments, and small ligands (less than five atoms). There are 13 protein and 13 ligand atom types described in Table 2. The protein atom types are assigned according to the atom position in the residue (for example, OD1 and OD2 atoms of aspartate

**Table 1.** Testing Set Complexes and Their Binding Constants[a]

| no. | PDB code | log $K_d$ | ligand name − protein name | no. | PDB code | log $K_d$ | ligand name − protein name |
|---|---|---|---|---|---|---|---|
| Subset 1: Aspartic Proteases (18 Complexes) | | | | | | | |
| 1 | 1aaq | −8.41 | hydroxyethylene − HIV protease | 10 | 1hvr | −9.52 | XK263 − HIV protease |
| 2 | 1hbv | −6.37 | SB203238 − HIV protease | 11 | 4hvp | −6.11 | MVT-101 − HIV protease |
| 3 | 1hpv | −9.23 | VX-478 − HIV protease | 12 | 4phv | −9.15[b] | L-700,417 − HIV protease |
| 4 | 1htf | −8.10 | GR126045 − HIV protease | 13 | 5hvp | −7.71 | acetyl-pepstatin − HIV protease |
| 5 | 1htg | −9.69 | GR137615 − HIV protease | 14 | 7hvp | −9.63 | JG-365 − HIV protease |
| 6 | 1hvi | −10.08 | A-77003 − HIV protease | 15 | 1apt | −9.41 | pepstatin analogue − penicillopepsin |
| 7 | 1hvj | −10.46 | A-78791 − HIV protease | 16 | 1apu | −7.71 | pepstatin analogue − penicillopepsin |
| 8 | 1hvk | −10.12 | A-76928 − HIV protease | 17 | 1ppk | −7.66 | phospho analogue − penicillopepsin |
| 9 | 1hvl | −9.01 | A-76889 − HIV protease | 18 | 1lyb | −11.43 | pepstatin − cathepsin D |
| Subset 2: Serine Proteases (20 Complexes) | | | | | | | |
| 19 | 1ppc | −6.46 | NAPAP − trypsin | 29 | 5lpr | −6.57[c] | AAPA-boronic acid − álytic protease (ALP) |
| 20 | 1pph | −6.23 | 3-TAPAP − trypsin | 30 | 6lpr | −7.30[c] | AAP-norleucineboronic acid − ALP |
| 21 | 1tng | −2.94 | aminomethylcyclohexane − trypsin | 31 | 7lpr | −7.18[c] | AAPL-boronic acid − ALP(M213A) |
| 22 | 1tnh | −3.37 | 4-fluorobenzylamine − trypsin | 32 | 8lpr | −6.62[c] | AAPF-boronic acid − ALP |
| 23 | 1tni | −1.70 | 4-phenylbutylamine − trypsin | 33 | 9lpr | −5.70[c] | AAPL-boronic acid − ALP |
| 24 | 1tnj | −1.96 | 2-phenylethylamine − trypsin | 34 | 3lpr | −9.59[c] | AAP-norleucineboronic acid − ALP(M192A) |
| 25 | 1tnk | −1.49 | 3-phenylpropylamine − trypsin | 35 | 1etr | −7.41 | MQPA − trombin |
| 26 | 1tnl | −1.88 | *t*−2-phenylcyclopropylamine − trypsin | 36 | 1ets | −8.53 | NAPAP − trombin |
| 27 | 3ptb | −4.74 | benzamidine − trypsin | 37 | 1ett | −6.19 | 4-TAPAP − trombin |
| 28 | 1bra | −1.83 | benzamidine − trypsin mutant | 38 | 1tmt | −6.24 | D-Phe-Pro-Arg − trombin |
| Subset 3: Metalloproteases (22 Complexes) | | | | | | | |
| 39 | 1tlp | −7.56 | phosphoramidon − thermolysin | 50 | 1mmr | −5.89[d] | sulfodiiminie inhibitor − MMP7 |
| 40 | 1tmn | −7.31 | peptidomimetic − thermolysin | 51 | 1jao | −5.92[d] | 3-mercapto-2-benzylpropanoyl-AG − MMP8 |
| 41 | 2tmn | −5.89 | P-Leu-NH$_2$ − thermolysin | 52 | 1mmb | −9.22[d] | hydroxamate inhibitor − MMP8 |
| 42 | 3tmn | −5.91 | VW − thermolysin | 53 | 1tlc | −8.05[d] | 1843U89 − MMP1 |
| 43 | 4tln | −3.72 | Leu-NHOH − thermolysin | 54 | 1mnc | −8.70 | hydroxamate inhibitor − MMP8 |
| 44 | 4tmn | −10.20 | ZFP(O)LA − thermolysin | 55 | 1cbx | −6.35 | L-benzylsuccinate − carboxypeptidase |
| 45 | 5tln | −6.37 | nitroanilide − thermolysin | 56 | 3cpa | −3.89 | GY − carboxypeptidase |
| 46 | 5tmn | −8.05 | thorphan − thermolysin | 57 | 6cpa | −11.53 | ZAAP(O)F − carboxypeptidase |
| 47 | 6tmn | −5.05 | ZGP(O)LL − thermolysin | 58 | 7cpa | −13.97 | BZ-FVP(O)F − carboxypeptidase |
| 48 | 1mmp | −6.24[d] | carbohylate inhibitor − MMP7 | 59 | 8cpa | −9.16 | BZ-AGP(O)F − carboxypeptidase |
| 49 | 1mmq | −9.00[d] | hydrohamate inhibitor − MMP7 | 60 | 2xis | −5.83 | xylitol − xylose isomerase |
| Subset 4: Human Carbonic Anhydrase II (HCA, 19 Complexes) | | | | | | | |
| 61 | ca_R | −10.52[e] | R-tby-indole − HCA | 71 | 1bnu | −9.70[i] | AL5300 − HCA |
| 62 | ca_S | −9.64[e] | S-tby-indole − HCA | 72 | 1bnt | −9.80[i] | AL5424 − HCA |
| 63 | 1am6 | −4.33[f] | methylhydroxamate − HCA | 73 | 1bnq | −9.49[i] | AL4623 − HCA |
| 64 | 1bcd | −3.9[g] | methylsulfonamide − HCA | 74 | 1a42 | −9.89[i] | brinzolamide − HCA |
| 65 | 1cil | 9.43[h] | ETS − HCA | 75 | 1bnn | −10.0[i] | AL7182 − HCA |
| 66 | 1cim | −8.82[h] | PTS − HCA | 76 | 1bnm | −10.0[i] | AL7089 − HCA |
| 67 | 1cin | −8.73[h] | MTS − HCA | 77 | 1bnv | −8.77[i] | AL7099 − HCA |
| 68 | 1bn1 | −9.34[i] | AL5917 − HCA | 78 | 1bn3 | −9.89[i] | AL6528 − HCA |
| 69 | 1bn4 | −9.31[i] | AL5927 − HCA | 79 | ca_F | −8.77[j] | inhibitor 1 − HCA |
| 70 | 1bnw | −9.08[i] | AL5415 − HCA | | | | |
| Subset 5: Sugar-Binding Proteins (14 Complexes) | | | | | | | |
| 80 | 1abe | −7.03 | L-arabinose − arabinose binding protein | 87 | 8abp | −8.01 | D-galactose − ABP(M108L) |
| 81 | 1abf | −5.43 | D-fucose − arabinose binding protein | 88 | 9abp | −8.01 | D-galactose − ABP(P254G) |
| 82 | 5abp | −6.65 | D-galactose − arabinose binding protein | 89 | 1nsd | −5.31 | DANA − neuraminidase |
| 83 | 1apb | −5.83 | D-fucose − ABP(P254G) | 90 | 1dog | −4.02 | 1-deoxynojirimycin − glucoamylase |
| 84 | 1bap | −6.87 | L-arabinose − ABP(P254G) | 91 | 1mfe | −5.32 | D-gal-D-abe-D-man − immunoglobulin |
| 85 | 6abp | −6.37 | L-arabinose − ABP(M108L) | 92 | 2gbp | −7.60 | D-glucose − glucose binding protein |
| 86 | 7abp | −6.47 | D-fucose − ABP(M108L) | 93 | 5cna | −2.00 | A-O1-methyl-mannose − concanavilin |
| Subset 6: Endothiapepsin (11 Complexes) | | | | | | | |
| 94 | 1eed | −4.80 | PD125754 − endothiapepsin | 100 | 2er9 | −7.81 | L363,564 − endothiapepsin |
| 95 | 1epo | −7.96 | CP-81,282 − endothiapepsin | 101 | 3er3 | −7.11 | CP-71,362 − endothiapepsin |
| 96 | 1epp | −7.17 | PD-130,693 − endothiapepsin | 102 | 4er1 | −6.63 | PD125967 − endothiapepsin |
| 97 | 2er0 | −6.40[b] | L-364,099 − endothiapepsin | 103 | 4er4 | −6.80 | H-142 − endothiapepsin |
| 98 | 2er6 | −7.22 | H-256 − endothiapepsin | 104 | 5er2 | −6.58 | CP-69,799 − endothiapepsin |
| 99 | 2er7 | −9.02 | H-261 − endothiapepsin | | | | |
| Subset 7: Purine Nucleoside Phosphatase (PNP, 5 Complexes)[k] | | | | | | | |
| 105 | pnp1 | −8.96 | inhibitor 1 − PNP | 108 | pnp4 | −5.85 | inhibitor 4 − PNP |
| 106 | pnp2 | −6.22 | inhibitor 2 − PNP | 109 | pnp5 | −5.96 | inhibitor 5 − PNP |
| 107 | pnp3 | −7.16 | inhibitor 3 − PNP | | | | |
| Subset 8: Other Proteins (10 Complexes) | | | | | | | |
| 110 | 1adb | −8.41 | CNAD − alcohol dehydrogenase | 115 | 1rbp | −6.72 | retinol − retinol-binding protein |
| 111 | 1ebg | −10.83 | phosphonoacetohydroxamate − enolase | 116 | 2cgr | −7.28 | GAS − immunoglobulin |
| 112 | 1fkf | −9.70 | FK-506 − FKPB | 117 | 2ifb | −5.43 | palmitic acid − fatty acid binding protein |
| 113 | 1hsl | −7.31 | histidine − histidine-binding protein | 118 | 2ypi | −4.82 | 2-phosphoglycolate − TP isomerase |
| 114 | 1pgp | −5.70 | 6-phosphogluconic acid − 6-PGDH | 119 | 4dfr | −9.71 | methotrexate − DHFR |

[a] Ref 14 unless otherwise noted. [b] Ref 13. [c] Ref 33. [d] Ref 34. [e] Ref 31; *R* and *S* stereoisomers of the inhibitor reported in this reference. [f] Ref 35. [g] Ref 36; crystal structure of $H_2NSO_2CF_3$ inhibitor was used for analysis of $H_2NSO_2CH_3$ assuming the same binding mode. [h] Ref 37. [i] Ref 38. [j] The crystal structure and binding constant were provided by Prof. D. W. Christianson. [k] The crystal structures and binding constants were provided by Prof. S. Elick.

are classified as charged oxygens; the OD atom of asparagine is classified as carbonyl oxygen). In ligands, the atom types are assigned by a separate procedure. The ligand part is extracted from the PDB file, and the hydrogen atoms are added

**Table 2.** SMoG Atom Types

| | |
|---|---|
| | Proteins |
| C3 | nonpolar $sp^3$ carbon (e.g., ALA C$\beta$) |
| C2 | nonpolar $sp^2$ carbon (e.g., PHE CG) |
| CA | C$\alpha$ mainchain carbon |
| CC | carbonyl and guanidinium $sp^2$ carbon (e.g., mainchain, ARG CZ, ASN CG) |
| CP | other polar (i.e., connected to at least one oxygen or nitrogen) carbon (e.g., TYR CZ) |
| OD | hydrogen bond donor oxygen (e.g., TYR OH) |
| OC | charged oxygen (e.g., ASP OD) |
| OB | carbonyl oxygen (e.g., mainchain, ASN OD) |
| ND | hydrogen bond donor nitrogen (e.g., TRP NZ) |
| NC | charged nitrogen (e.g., ARG NE, NH, HIS ND,NE) |
| NM | mainchain nitrogen |
| SP | sulfur |
| M | any metal ion |
| | Ligands |
| C3 | nonpolar $sp^3$ carbon |
| C2 | nonpolar $sp^2$ carbon |
| CC | carbonyl and guanidino $sp^2$ carbon |
| CP | other polar (i.e., connected to at least one oxygen or nitrogen) carbon |
| OD | hydrogen bond donor oxygen (e.g., hydroxy) |
| OC | charged oxygen (e.g., carboxylate, phosphate) |
| OB | carbonyl oxygen (e.g., amide, keto) |
| OA | hydrogen bond acceptor oxygen (e.g., ether) |
| ND | hydrogen bond donor nitrogen (e.g., secondary amine, pyrrol) |
| NC | charged nitrogen (e.g., primary amine) |
| NA | hydrogen bond acceptor nitrogen (e.g., pyridine) |
| NM | amide nitrogen (in peptide ligands) |
| SL | sulfur and phosphorus |

by Babel. Hydrogens are necessary for determining the valences of certain ligand atoms to assign their atom types (for example, donor $sp^3$ nitrogen has the valence of three, whereas charged $sp^3$ nitrogen has the valence of four). No filtering of the training set according to the resolution and the homology of proteins or ligands is performed. It is also important to note that the training set includes 51 complexes from the testing set. In separate control runs, we excluded these complexes from the training set and reproduced the correlation coefficients for all subsets in all tests (data not shown).

## Results and Discussion

**SMoG2001 Has Improved Predictive Power over SMoG96.** Figure 2A shows the scatter plot of experimental binding constants vs SMoG96 scores (eqs 1−4) for 119 ligands of the testing set, together with the linear fits, correlation coefficients, and standard deviations for subsets of individual protein classes. The correlation coefficient for the whole set is low (0.304); for individual classes, it exceeds 0.5 only for serine proteases and carbonic anhydrase subsets. For sugar-binding complexes, the correlation coefficient is negative. The linear fits for the subsets are distant. It is evident that SMoG96 performs poorly even in predicting binding affinities of ligands within subsets, let alone the relative position of the subsets with respect to one another.

Prediction of binding affinities using the SMoG2001 scoring function (eqs 5−7) is considerably improved (Figure 2B). The overall correlation coefficient increases from 0.304 (SMoG96) to 0.435 (SMoG2001). The correlation coefficients are also higher for the majority of the subsets. In particular, substantial improvement is achieved for aspartic proteases (0.16−0.62), sugar-binding proteins (−0.50 to 0.47), and the class of other proteins (0.19−0.78). A slight decrease in correlation coefficient (by about 0.1) is observed for carbonic anhydrase and endothiapepsin. The majority of the linear fits to the data are parallel and close to each other.

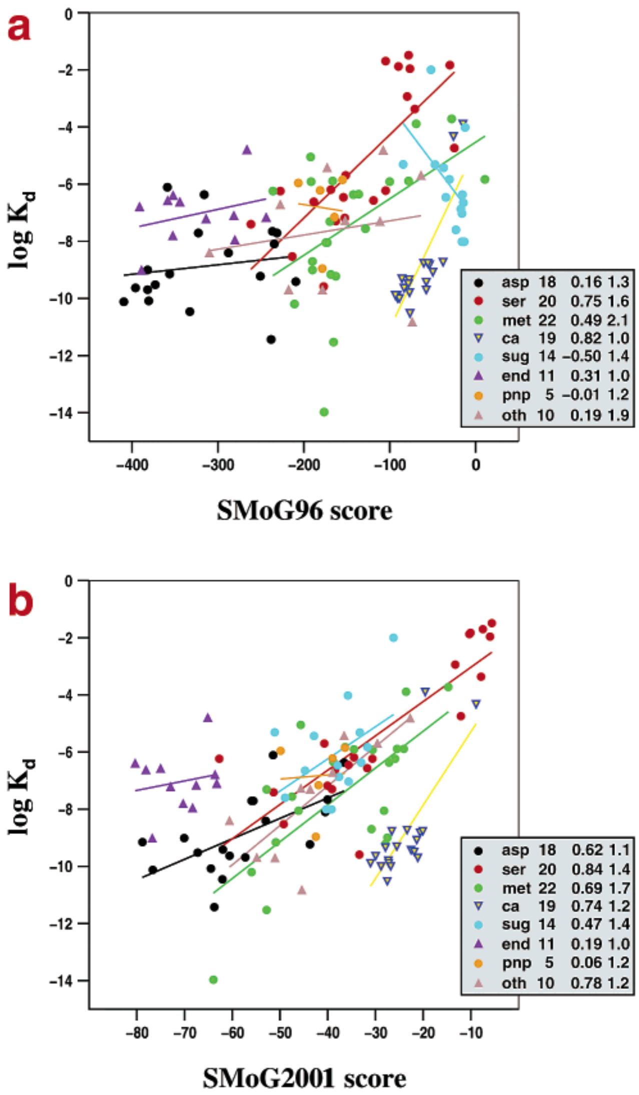

**Figure 2.** Plot of (a) SMoG96 scores and (b) SMoG2001 scores of 119 protein−ligand complexes of the testing set vs their experimental binding constants (log $K_d$). Symbols correspond to eight subsets of structurally related proteins (asp, aspartic proteases; ser, serine proteases; met, metalloproteases; ca, carbonic anhydrase; sug, sugar-binding proteins; end, endothiapepsin; pnp, purine nucleoside phosphorylase; oth, other proteins) and are given on the insert, gray background. Linear fits to data are shown as lines of the same color as symbols of subsets. The insert also contains information for each subset in the following order: number of complexes, correlation coefficient, and standard deviation (The standard deviations were computed as $\sigma = \langle\sqrt{(\log K_d - (AF+B))^2}\rangle$ where log $K_d$ is the experimental binding constant, $AF + B$ is the equation of the line fitted to the data and $F$ is SMoG score.) from linear fit in units of log $K_d$.

Binding constants of the majority of the complexes are predicted well, with the standard deviation from the linear fit of 2.07 units of log $K_d$. There are, however, two subsets that are obvious outliers in the graph: SMoG2001 significantly overestimates binding of the endothiapepsin ligands (by 4 orders of magnitude) and underestimates the affinities of the carbonic anhydrase ligands (also by about 4 orders of magnitude). We believe that such inaccuracy results from inadequacy of a knowledge-based approach (in particular, SMoG2001) in accounting for strong interactions between ligand atoms and protein metal ions (carbonic anhydrase) and conformational entropy loss upon binding of large flexible ligands (endothiapepsin) as will be discussed below. When these subsets are not included into our analysis, the correlation coefficient becomes 0.77 (as compared to 0.53 in SMoG96) and the standard deviation becomes 1.52 units of log $K_d$ (as compared to 2.01 in SMoG96). The largest improvement of the SMoG2001 scoring function over SMoG96 results from better prediction of very high-affinity (log $K_d$ less than −11) as well as very low-affinity (log $K_d$ greater than −6) binders.

**Factors Responsible for the Improvement.** We believe that the improvement in the performance of the SMoG2001 over SMoG96 scoring function is a consequence of more accurate treatment of the reference state. In our previous work,[25] we compared the performance of the two derivation methods (SMoG96 and SMoG2001) in recovering back the true potential for the toy database of complexes using a simplified atom-typing scheme (five protein and five ligand atom types). In that work, SMoG96 failed to derive accurate potential and yielded a low correlation coefficient of 0.2 that did not change upon increase of the size of the training set (Figure 1A in ref 25). In contrast, the correlation coefficient calculated by SMoG2001 (with $\alpha = \beta = 1.0$ in the function for the reference state probabilities, eq 7) raised with the increase of the size of the training database reaching a high value of 0.8 at 100 complexes. The correlation coefficient calculated by this function, however, was low when $\alpha = \beta < 0.6$ was used (Figure 2A in ref 25). We therefore suggested that the main reason for the contrasting performance of SMoG2001 and SMoG96 methods for the toy database was the proper scaling of probabilities of contacts as $N(\sigma_p)N(\sigma_l)$ in SMoG2001 (the use of $\alpha = \beta = 1.0$ in the function for the reference state probabilities, eq 7) as opposed to $[N(\sigma_p)N(\sigma_l)]^{1/2}$ in SMoG96 (eq 3).

We wanted to investigate whether the scaling with the size of the training set and dependence on α and $\beta$ that we observed for the toy database have similar behavior for real protein−ligand complexes. The robust force field must be stable with respect to the training set size and composition; i.e., the parameters and, therefore, the scores of the ligands must converge to a single solution, which would not depend on the number of complexes in the training set from which the potential is derived. We also investigated how the definition of the reference state (values of α and $\beta$ in eq 7) would affect the correlation coefficient.

Figure 3 shows the dependencies of (i) the correlation coefficient between computed by SMoG2001 and experimental binding free energies of the testing set on the size of the training set (Figure 3A) and (ii) the absolute "statistical errors"[27] of the $F(r,\sigma_p,\sigma_l)$ values (defined as $\Delta F(r,\sigma_p,\sigma_l) = \sum_{\sigma_p}\sum_{\sigma_l}|F^{725}(r,\sigma_p,\sigma_l) - F^n(r,\sigma_p,\sigma_l)|$, where $F^{725}(r,\sigma_p,\sigma_l)$ is the parameter computed from the full training set of 725 complexes and $F^n(r,\sigma_p,\sigma_l)$ is that computed from the subset of randomly chosen $n$ complexes, $0 < n < 725$) for (i) all, (ii) the 10 most frequent, and (iii) the 10 least frequent contacts on the size of the training set (Figure 3B). We also note that 10 least frequent contacts on this graph are repulsive ones, i.e., those for which $F(r,\sigma_p,\sigma_l) > 0$.

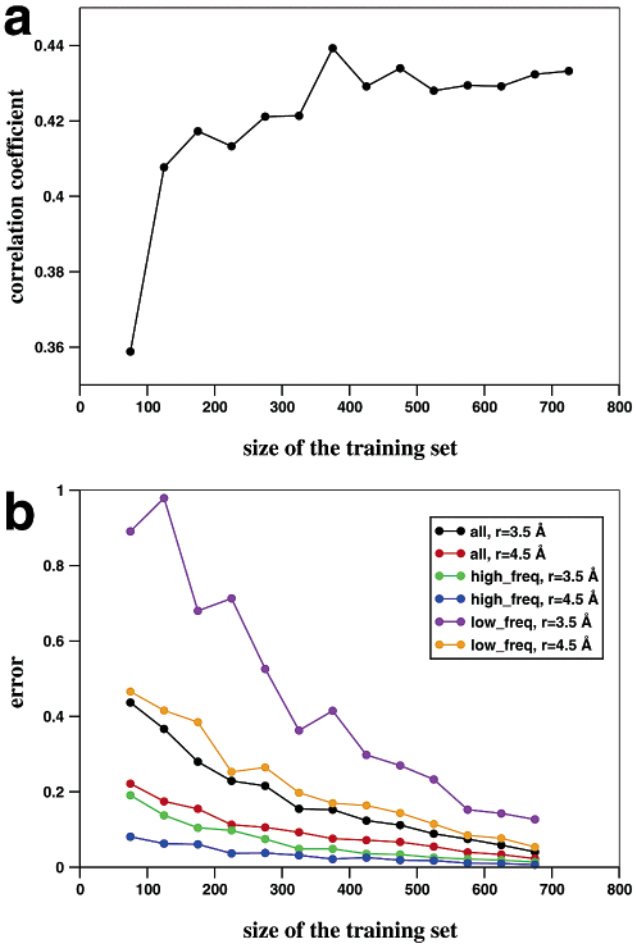

**Figure 3.** (a) Correlation coefficient of SMoG2001 scores vs experimental binding free energies of testing set vs size of the training set; that is, number of complexes from which the potential is derived. (b) Absolute "statistical error" (difference between the parameters computed from full training set and those computed from part of this set) as a function of size of the training set. The insert corresponds to the following: all, all 169 parameters; high freq, 10 parameters for most frequent contacts in full training set; low freq, 10 parameters for least frequent contacts in full training set; $r = 3.5$ Å, outer radius of the first distance interval; and $r = 4.5$ Å, outer radius of the second distance interval. Each graph represents an average over 10 separate runs. In these calculations, $\alpha = \beta = 0.9$, $r_{c1} = 3.5$ Å, and $r_{c2} = 4.5$ Å are used.

Initially (below 400 complexes), the correlation coefficient increases upon addition of complexes to the training database. For small training sets, occurrences of contacts between certain atom types are low; therefore, statistical errors in the $F(r,\sigma_p,\sigma_l)$ values are large (for example, for the interval 0−3.5 Å, the absolute convergence error of 10 least frequent contacts is ∼1, or the relative convergence error ∼100%, when the potential is derived from 125 complexes, Figure 3B). As the number of complexes in the training set increases, the statistical errors in the frequencies of contacts decrease for all atom types, and most potential parameters converge. The correlation coefficient reaches the constant value of around 0.43 after 500 complexes. Our complete training database is large enough to give stable a correlation coefficient for the testing set; however, the parameter values of the "repulsive con-

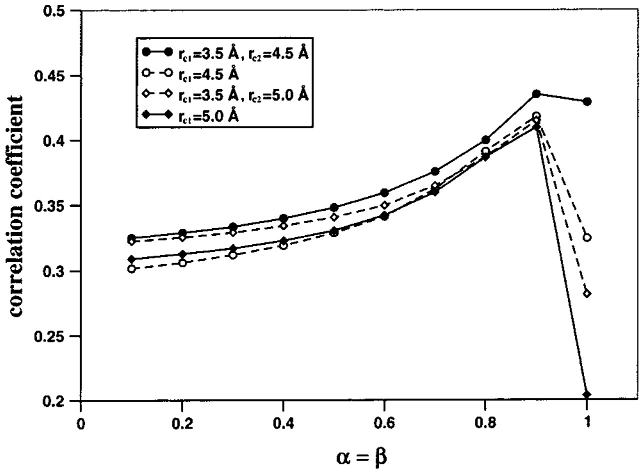

**Figure 4.** Dependence of the correlation coefficient in the testing set on α and $\beta$, eq 7. Here, α = $\beta$. Data for four distance cutoffs are shown; symbols corresponding to the outer radii of the distance intervals are shown on the insert.

tacts" (those that are rarely observed in the database) are still not statistically significant.

Figure 4 shows the dependence of the correlation coefficient on the values of exponents α and $\beta$ used in the definition of the reference state (eq 7) for several cutoff radii. The data corresponding to two single cutoffs (at 4.5 and 5.0 Å) as well as two double cutoffs (first at 3.5 Å and second at 4.5 and 5.0 Å) are shown. For all cutoff schemes, the correlation coefficient gradually increases with the increase of α and $\beta$, reaching a maximum of 0.43 at α = $\beta$ = 0.9. For α = $\beta$ = 1.0, the correlation coefficient decreases sharply. The value of the third free parameter, $F_0(r)$, does not affect the correlation coefficient for all values of α, $\beta$, and distance cutoff radii (data not shown).

We demonstrate in Figures 3 and 4 that when applied to the real protein−ligand complexes, the SMoG2001 scoring function manifests the same trends that were observed for the toy complexes: the potential is statistically robust, and the dependence of correlation coefficient on the parameters α and $\beta$ is similar. On the other hand, the correlation coefficient decreases at α = $\beta$ = 1.0 for real complexes, whereas it is maximal for the toy complexes for these values. We performed calculations for the toy database with 13 ligand atom types and 13 protein atom types, and the correlation coefficient decreased for α = $\beta$ = 1.0 as in the case of real complexes (data not shown). Therefore, this decrease occurs when more atom types are used and is similar for toy and real complexes.

In the previous work, we showed that the use of α = $\beta$ = 0.5 overestimates the strength of the atom−atom contacts involving at least one carbon atom. A scheme where such values of α and $\beta$ are used would, therefore, overestimate the scores of the complexes in which hydrophobic interactions dominate and underestimate those of the complexes in which polar and charged interactions dominate. We note that the scores of sugar-binding protein complexes that are stabilized mainly by hydrogen bonds are higher than those of the complexes with similar binding constants for the SMoG96 potential function and, therefore, overestimated; such scores are similar in the SMoG2001 function. SMoG2001 estimates correctly the affinities of ligands of sugar-binding proteins generally as well as yields a much higher correlation coefficient within the subset of sugar-binding proteins and other proteins (where the majority of the complexes are stabilized by polar and charged interactions). These results show that SMoG2001 might be more accurate than SMoG96 in estimating polar and hydrogen-bonding interactions.

**Table 3.** Correlation Coefficients Computed Using Nonspecific Contact Pairwise Scoring Function ($F(r, \sigma_p, \sigma_l) = -1$) and SMoG2001

| subset | no. of complexes | SMoG2001 | nonspecific function |
|---|---|---|---|
| aspartic proteases | 18 | 0.622 | 0.341 |
| serine proteases | 20 | 0.835 | 0.614 |
| metalloproteases | 22 | 0.688 | 0.729 |
| carbonic anhydrase | 19 | 0.744 | 0.800 |
| sugar-binding proteins | 14 | 0.471 | 0.046 |
| endothiapepsin | 11 | 0.187 | 0.598 |
| pnp | 5 | 0.056 | 0.267 |
| other proteins | 10 | 0.778 | 0.432 |
| combined | 119 | 0.435 | 0.356 |

We note that the correlation coefficient does not change significantly when two distance intervals are used; the best two step potential ($r_{c1}$ = 3.5 Å, $r_{c2}$ = 4.5 Å) yields only a slightly better correlation coefficient (by about 0.02) than the best single-step function ($r_c$ = 4.5 Å). In principle, use of distance dependence should increase the precision of the scoring function because of the lesser degree of averaging of the $F(r,\sigma_p,\sigma_l)$ values. In particular, effective distance for hydrogen bonding (3.5 Å) is smaller than for hydrophobic (5 Å) and polar contacts (>5 Å), and $F(r,\sigma_p,\sigma_l)$ values of atom types participating in the hydrogen bonding should be more precise if we include $r_c$ = 3.5 Å. Figure 4 shows that incorporation of this distance interval does not influence the results significantly. Use of multiple intervals is less desirable because statistical errors in frequencies of contacts increase. Therefore, given slight improvement using two cutoff radii, we do not further increase the number of distance bins and use two intervals in our potential function.

We also investigate the effect of the specificity of the scoring function to different atom types (i.e., $F(r,\sigma_p,\sigma_l)$ values being specific to $\sigma_p$ and $\sigma_l$) on the correlation coefficients. To this end, we compare the correlation coefficients computed using the SMoG2001 function with those computed using a nonspecific function, that is, a pairwise contact potential with the same definition of atom−atom contacts (contact if distance between atoms is less than 4.5 Å; no contact otherwise) but with all $F(r,\sigma_p,\sigma_l) = -1$. The results are given in Table 3: SMoG2001 gives significantly higher correlation coefficients for aspartic proteases, serine proteases, sugar-binding proteins, and other proteins. We note that in contrast, SMoG96 performs worse than the nonspecific function for aspartic proteases, sugar-binding proteins, and other proteins; this is another demonstration of improvement of SMoG2001 over SMoG96. For metalloproteases and carbonic anhydrase, SMoG2001 performs slightly worse than the nonspecific function. For PNP and endothiapepsin, nonspecific function actually performs better. Such complexes are either rich in contacts whose $F(r,\sigma_p,\sigma_l)$ values are not accurate or they have the features that are not accounted for in KBP.

**Table 4.** Correlation Coefficients and Standard Deviations Computed for the Subsets of Proteins Comprising the Testing Set of 77 Complexes Used in Validation of PMF Scoring Function[a]

| | SMoG2001 | | PMF | | SCORE1 | | DrugScore | |
|---|---|---|---|---|---|---|---|---|
| subset | R | SD | R | SD | R | SD | R | SD |
| serine proteases (16) | 0.90 | 1.09 | 0.87 | 0.96 | 0.76 | 1.39 | 0.86 | 0.95 |
| metalloproteases (15) | 0.80 | 1.62 | 0.58 | 2.31 | 0.41 | 3.27 | 0.70 | 1.53 |
| sugar-binding proteins[b] (18) | 0.24 | 0.80 | 0.48 | 0.86 | 0.00 | 69.7 | 0.22 | 0.75 |
| endothiapepsin (11) | 0.18 | 1.00 | 0.22 | 1.89 | 0.39 | 1.26 | 0.30 | 0.94 |
| other proteins (17) | 0.71 | 1.63 | 0.69 | 1.56 | 0.53 | 2.21 | 0.43 | 1.85 |
| combined (77) | 0.68 | 1.69 | 0.61 | 1.84 | 0.30 | 3.47 | N/A[c] | |

[a] The data are reported for SMoG2001, PMF, SCORE1 and Drugscore scoring functions. [b] For the complexes of sugar-binding proteins, there are nine different PDB structures containing both α and $\beta$ forms of sugar ligands for which the scores were computed separately; corresponding binding constants were obtained for racemic mixtures of compounds and are therefore identical for both isomers of each complex. [c] Not available.

For these complexes, KBP does not predict binding affinities well.

Next, we compare SMoG2001 with other scoring functions widely used for predicting binding affinities. To this end, we decided to use the available correlation data reported by Muegge and Martin[18] and Gohlke et al.[21] The data were reported for the original versions of three scoring functions—PMF,[18] a KBP developed by Muegge and Martin; DrugScore,[20] another KBP developed by Gohlke et al.; and SCORE1(LUDI),[11] an empirical force field developed by Böhm; we note that an improved scoring function by Böhm (SCORE2)[12] as well as more studies using various modifications of PMF[28,29] have been published. PMF and DrugScore use slightly different equations to relate the contact frequencies in the database to free energies and incorporate distance dependence; that is, the statistics were collected over small bins of 0.2 (PMF) or 0.1 Å (DrugScore); PMF also uses additional ligand volume correction factor, and DrugScore incorporates a term accounting for solvent accessibility.

Table 4 and Figure 5 show the correlations and standard deviations in the testing set of 77 complexes and five subsets corresponding to proteins of different types used in the original validation of PMF function. Overall, the correlation coefficient and standard deviation computed by SMoG2001 (0.68, 1.69) are slightly better than those computed by PMF (0.61, 1.84). For different subsets, the correlation coefficients for serine proteases, endothiapepsin, and a class of other proteins are similar. SMoG2001 performs better than PMF for metalloproteases. In the case of the complexes of sugar-binding proteins involving many hydrogen bonds, the PMF function reproduces binding affinities better than SMoG2001 for the same set of complexes that was used in the PMF study. For the extended set of sugar-binding complexes (used in Figure 2), however, SMoG2001 gives a higher correlation coefficient. We note that the outliers are the same in both approaches: 1tmt is the outlier in serine proteases, and 1mnc is the outlier for metalloproteases; binding affinities of all endothiapepsin ligands are overestimated by PMF as well.

As compared to DrugScore, SMoG2001 shows slightly better correlation coefficients for all subsets except endothiapepsin. For this subset, DrugScore performs better than the two other KBP. SMoG2001 reproduces the binding affinities of the ligands of the class of other proteins better than DrugScore. The standard deviations from linear fits obtained by DrugScore are slightly smaller than those obtained by SMoG2001 for all classes except other proteins.

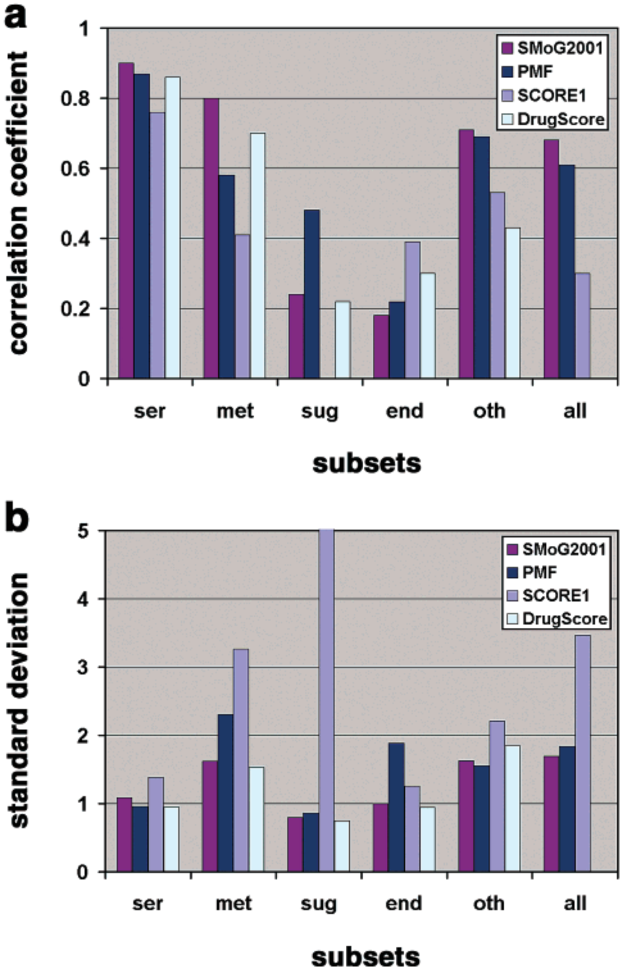

**Figure 5.** Comparison between SMoG2001, PMF, SCORE1, and DrugScore: (a) correlation coefficients and (b) standard deviations in units of log $K_d$. The corresponding data are given in Table 4.

SMoG2001 is a simpler scoring function than PMF and DrugScore. It does not include the distance dependence, it has fewer atom types, and it does not use extra terms (such as ligand volume correction factor in PMF or solvent accessibility term in DrugScore). Also, although the authors of the PMF function argue that the use of a large cutoff (12 Å) for the reference state improves the accuracy of the force field by implicitly including solvation effects, our results show that the optimal cutoff is 4.5 Å and the correlation coefficients decrease upon increasing this parameter. We hypothesize that the use of a different definition of the reference state (proportional to the number of contacts along the protein−ligand interface and excluding the atoms deep inside the protein, which may add noise to the statistics) may be responsible for such difference in the behavior of our scoring function and PMF. We also believe that it is not the spatial resolution but the definition of the reference state that is crucial to the good predictive power of a knowledge-based function.

SMoG2001 shows better correlation than SCORE1 for all subsets except endothiapepsin. For the whole set, the correlation coefficient and standard deviation computed by SMoG (0.68, 1.69) are greater than those computed by SCORE1 (0.30, 3.47). For long flexible endothiapepsin ligands, empirical function SCORE1 scores better than all three KBP.

**SMoG2001 Mishandles Interactions with Significant Quantum Effects.** Metal binding has a different nature than the other interactions present in protein−ligand complexes. The contacts between metal and ligand atoms are shorter (2 Å), stronger, and more geometry-dependent than other contacts. Interactions between metal ions and ligand groups can be accurately described only by quantum mechanical methods. Metal−ligand interactions dominate the binding free energy (for example, 1BCD ligand of carbonic anhydrase consisting of zinc-bound sulfonamide and one methyl group binds with log $K_d = -3.9$ or $\Delta G = -5.5$ kcal/mol, whereas all other sulfonamide ligands of this protein included in our testing set have log $K_d < -9$, or $\Delta G < -12.2$ kcal/mol; therefore, the zinc−sulfonamide interaction is about 40% of binding free energy of carbonic anhydrase inhibitors). Given the presence of such unusual interaction, it is encouraging to observe that the binding affinities of many metalloprotease complexes are calculated with the accuracy comparable to the other protein classes. There are, however, several exceptions. In particular, SMoG2001 underestimates the affinities of all carbonic anhydrase ligands; we do not know why the interaction with zinc is unusually strong for the ligands of this protein as compared to the other metalloproteases; so that our scoring function underestimates it.

In Figure 6, 41 complexes of the metalloproteases are subdivided to the sets according to metal-binding functional groups of the ligand. Generally, the scoring function can accurately predict the binding affinities within the sets. However, the prediction of the sets relative to each other is poor as the linear fits to the subset data differ significantly in slopes and intercepts. We speculate that this difference is due to the inability of the method to account for quantum forces that drive the interaction between metal and ligand and are different for various metal-chelating groups. Incorporation of a separate term dependent on the strength of metal binding for a particular functional group into a knowledge-based scoring function should improve prediction of complexes in which metal−ligand interactions are present.

**Additional Term for Flexible Bonds.** Scatter plot in Figure 2 shows that SMoG2001 overestimates the binding affinities of endothiapepsin ligands. All ligands of this aspartic protease are peptide chains (about six residues) with many flexible bonds (about 25); these bonds are frozen upon binding, resulting in the loss of the conformational entropy of the ligand. Experimental estimates of the loss of free energy per one flexible bond during binding are 0.4−0.9 kcal/mol.[30]

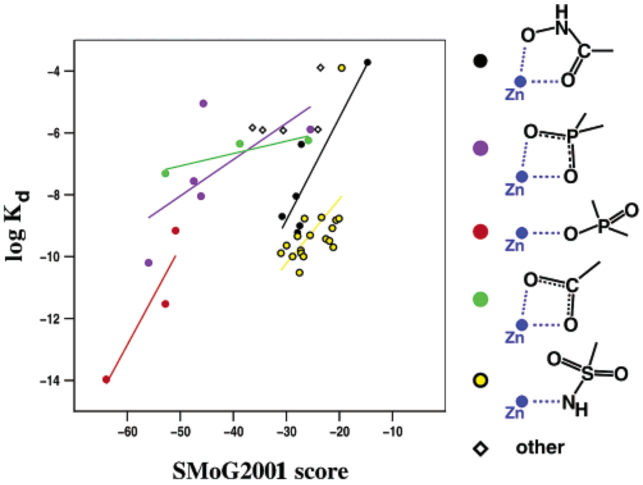

**Figure 6.** Plot of SMoG2001 scores vs experimental binding constants for metal-containing 22 metalloprotease complexes and 19 carbonic anhydrase complexes. Symbols correspond to subsets of the complexes with the same ligand functional group coordinating to the metal. The interactions between metal ions and ligand functional groups defining each subset are shown. Linear fits to data are indicated.

Because a knowledge-based method does not account for the loss of the conformational entropy upon binding, we do not expect it to correctly predict binding free energies for highly flexible ligands such as those of endothiapepsin. To better handle such ligands, we added a heuristic "entropic" term to the pure knowledge-based function. We define this term as follows.

We classify all ligand nonterminal $sp^3$−$sp^3$ and $sp^3$−$sp^2$ bonds as flexible bonds, which are frozen upon binding. Terminal bonds are presumed to rotate freely in both free and complexed ligands and therefore not included. Aliphatic ring bonds are not included. We divide the flexible -A($X_2$)-A($X_2$)- bonds into two categories: (i) bonds with unrestricted rotations, for which at least one atom A does not have any heavy atom X connected to it (such as -$CY_2$-$CH_2$-, -$CY_2$-O-; Y denotes any atom) and (ii) bonds with restricted rotations, in which both atoms A are connected to at least one nonhydrogen atom X (for example, -CNH-CCH-). The loss of conformational entropy associated with the second type of flexible bond is smaller than that of the first type since there are fewer sterically available states in free ligand and the degree of its freezing upon binding is smaller. With the inclusion of the entropy term, our scoring function becomes

$$F^{*} = F^{\text{kn\_based}} + F^{\text{rot1}}\, N^{\text{rot1}} + F^{\text{rot2}}\, N^{\text{rot2}} \qquad (8)$$

where $F^{*}$ is the binding free energy; $F^{\text{kn\_based}}$ is the part to the binding free energy calculated by knowledge-based formalism (eq 4); $F^{\text{rot1}}$ and $F^{\text{rot2}}$ are the parts of the binding free energy associated with the freezing of the flexible bonds of type 1 and 2; $N^{\text{rot1}}$ and $N^{\text{rot2}}$ are the numbers of the flexible bonds of type 1 and 2 in the ligand.

We find the values $F^{\text{rot1}}$ and $F^{\text{rot2}}$ by optimizing the correlation coefficient in the testing set; these optimal values are $F^{\text{rot1}} = 1.2$ and $F^{\text{rot2}} = -0.5$. The scatter plot of the experimental binding constants vs binding free

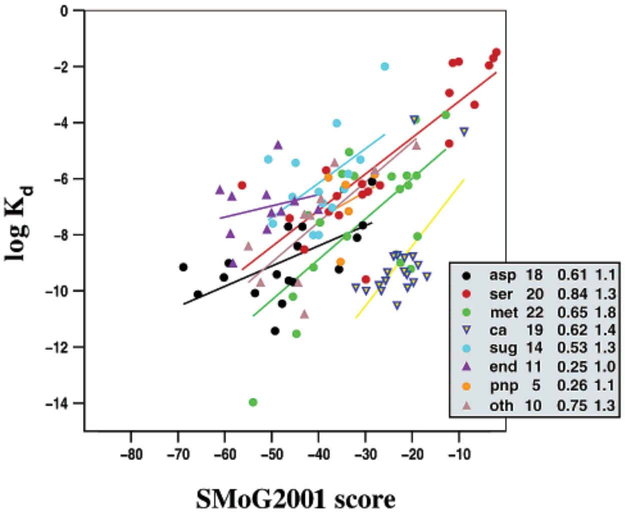

**Figure 7.** Plot of scores of protein−ligand complexes of the testing set computed by eq 8 (knowledge-based function combined with ligand conformational entropy term) vs their experimental binding constants (log $K_d$). Insert contains the same information as in Figure 2. Linear fits to data are shown as lines of the same color as symbols of subsets.

energies calculated using eq 8 is shown in Figure 7. No significant change in the correlation coefficient is observed either for the entire set or for the individual subsets with small ligands. On the other hand, we note that the binding of endothiapepsin ligands is predicted better, overestimated by only 2 orders of magnitude relative to the other subsets (as compared to 4 orders without a conformational entropy term, Figure 2). We suggest that the conformational entropy term should be included in such applications as design of large flexible ligands.

## Conclusions and Outlook

The definition of the reference state is a key factor in the performance of a knowledge-based potential. Using the redefined reference state in the SMoG2001 scoring function, we obtained a higher correlation coefficient (0.435) than that computed using SMoG96 (0.304) for a diverse set of 119 complexes as well as for the smaller subsets. A simple, coarse-grained, pairwise SMoG2001 potential with only 13 atom types and two distance intervals can calculate $K_d$ values of the majority of complexes with the accuracy of about 2 orders of magnitude, which is somewhat better than that of the other knowledge-based methods as well as empirical scoring functions (see Figure 5). Our method can be used in computational de novo lead generation projects, and we have recently demonstrated its success by designing two picomolar inhibitors for human carbonic anhydrase II using the SMoG2001 scoring function and CombiSMoG growth algorithm.[31]

The size of our training database is sufficient for constructing the force field, which gives stable correlation of the testing set. We found, however, that for less frequent contacts the statistical errors in the potential of mean force are still large at this size of training set. In future design applications, ligands will be generated in which such contacts will be more abundant than in testing set complexes used in our study. To evaluate the scores of these weak-binding, putative ligands accurately, all force field parameters, including repulsive ones, should have small errors. Therefore, to further improve the accuracy of a knowledge-based scoring function for design of new compounds, one should use a database that would include a larger number of poor binders, i.e., complexes with more abundant repulsive contacts.

Although a considerable amount of work has been already done on the development of KBPs, these potentials do not significantly differ in terms of accuracy.[18−21,28,29] We believe that there is little room for improvement in the ways in which the frequencies of contacts are converted to binding free energies. In accord with the previous suggestions,[18] our analysis shows that the performance of KBP cannot be improved by simply increasing the size of the training database or by varying its composition according to homology of the proteins; the database of ∼500 complexes is sufficient to produce statistically robust knowledge-based force fields. We hypothesize that the accuracy can be improved by optimizing the numbers and definitions of atom types. We plan to study these issues in the future. Also, it is possible that the incorporation of empirical terms not accounted for in a knowledge-based approach (for example, ligand strain energy) would lead to better scoring functions. We showed here, however, that the incorporation of the ligand conformational entropy term does not significantly improve the accuracy of SMoG2001 except for large peptide ligands.

We believe that main weaknesses of the KBP lie in estimation of directional (such as metal binding or hydrogen bonds), polar, and repulsive interactions. As for hydrogen bonds, incorporating statistics derived from the Cambridge Structural Database of the crystals of small organic molecules might be helpful; the methodology of such derivation can be found in our previous work.[32]

More accurate KBPs can be possibly constructed if one can better understand how the scoring function describes the relative contributions of interactions of various types (polar interactions, hydrogen bonds, etc.) in the total binding free energy of the complex. A small testing set of about 30−50 related compounds of known three-dimensional structures and binding affinities and of similar topologies of polar interactions or hydrogen bonds but different scaffolds should be helpful in understanding the energetics of these interactions. It would also be beneficial to include the series of “mutated” ligands (for example, an important hydrogen bond donor at a particular place replaced by a hydrogen bond acceptor, charged group, hydrophobic group, or simply removed). An interesting example to probe the metal−ligand interactions would involve the potent carbonic anhydrase inhibitors with other than a sulfonamide zinc-binding group or without this group altogether. The interactions investigated this way should not be energetically dominant; that is, there should be several types of interactions of comparable magnitude. These complexes can be included in the training database if necessary to improve the accuracy of the KBP.

We also think that the accuracy of KBP can be improved by incorporating in the training set more complexes that exhibit infrequent contacts (such as metalloprotein ligands or weakly bound complexes). For the weakly bound complexes, the ligands obtained from a design round using KBP and predicted to bind with low affinity can be experimentally characterized and included in the training set. By comparing the predicted and experimental structures and free energies of the designed ligands, one could possibly detect important aspects in which the accuracy of KBPs can be improved.

**Acknowledgment.** This work was supported by NIH (Grant no. RO1-52126). We thank Dr. Bartosz A. Grzybowski for helpful discussions and critically reading the manuscript and Jun Shimada for substantial help during all stages of this project.

## References

(1) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand−receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.

(2) Martin, Y. C. Challenges and prospects for computational aids to molecular diversity. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 159−172.

(3) Joseph-McCarthy, D. Computational approaches to structure-based ligand design. *Pharmacol. Ther.* **1999**, *84*, 179−191.

(4) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiner, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3585−3616.

(5) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force-field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5187.

(6) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(7) Halgren, T. A. Merck molecular force field: 1. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(8) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. Molecular mechanics−the MM3 force field for hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111*, 8551−8566.

(9) Kollman, P. Free energy calculations−applications to chemical and biological phenomena. *Chem. Rev.* **1993**, *7*, 2395−2417.

(10) Kollman, P. A. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc. Chem. Res.* **1996**, *29*, 461−469.

(11) Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(12) Böhm, H.-J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.

(13) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(14) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(15) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein−ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.

(16) Grzybowski, B. A.; Ishchenko, A. V.; Shimada, J.; Whitesides, G. M.; Shakhnovich, E. I. From knowledge-based potentials to combinatorial lead design in silico *Acc. Chem. Res.* **2002**, 5.

(17) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de novo design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.

(18) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(19) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP−potential of mean force describing protein−ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

(20) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(21) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein−ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.

(22) Wallqvist, A.; Jernigan, R. L.; Covell, D. G. A preference-based free energy parametrization of enzyme−inhibitor binding−applications to HIV-1 protease inhibitor design. *Protein Sci.* **1995**, *4*, 1881−1903.

(23) Verkhiver, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical free energy calculations of protein−ligand crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.* **1995**, *8*, 677−691.

(24) DeWitte, R. S.; Ishchenko, A. V.; Shakhnovich, E. I. SMoG: de novo design method based on simple, fast and accurate free energy estimates. 2. Case studies in molecular design. *J. Am. Chem. Soc.* **1997**, *119*, 4608−4617.

(25) Shimada, J.; Ishchenko, A. V.; Shakhnovich, E. I. Analysis of knowledge-based protein−ligand potentials using a self-consistent method. *Protein Sci.* **2000**, *9*, 765−775.

(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(27) Given our model, there are two sources of errors in the parameters: (i) that the training database is incomplete, i.e., the number of complexes is small, and (ii) that the database composition is biased towards complexes of particular contact patterns. The "statistical errors" described here estimate the first source of errors.

(28) Muegge, I. A knowledge-based scoring function for protein−ligand interactions: probing the reference state. *Perspect. Drug Discovery Des.* **2000**, *20*, 99−114.

(29) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.

(30) Searle, M. S.; Williams, D. H. The cost of conformational order−entropy changes in molecular associations. *J. Am. Chem. Soc.* **1992**, *114*, 10690−10697.

(31) Grzybowski, B. A.; Ishchenko, A. V.; Kim, C.-Y.; Topalov, G.; Chapman, R.; Christianson, D. W.; Whitesides, G. M.; Shakhnovich, E. I. Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1270−1273.

(32) Grzybowski, B. A.; Ishchenko, A. V.; DeWitte, R. S.; Whitesides, G. M.; Shakhnovich, E. I. Development of a knowledge-based potential for crystals of small organic molecules: calculation of energy surfaces for C═O...H−N hydrogen bonds. *J. Phys. Chem. B* **2000**, *104*, 7293−7298.

(33) Bone, R.; Fujishige, A.; Kettner, C. A.; Agard, D. A. Structural basis for broad specificity of alpha-lytic protease mutants. *Biochemistry* **1991**, *30*, 10388−10398.

(34) Babine, R. E.; Bender, S. L. Molecular recognition of protein−ligand complexes: applications to drug design. *Chem. Rev.* **1997**, *97*, 1359−1472.

(35) Scolnick, L. R.; Clements, A. M.; Liao, J.; Crenshaw, L.; Hellberg, M.; May, J.; Dean, T. R.; Christianson, D. W. Novel binding mode of hydroxamate inhibitors to human carbonic anhydrase II. *J. Am. Chem. Soc.* **1997**, *119*, 850−851.

(36) Maren, T. H.; Conroy, C. W. A new class of carbonic anhydrase inhibitor. *J. Biol. Chem.* **1993**, *268*, 26233−26239.

(37) Smith, G. M.; Alexander, R. S.; Christianson, D. W.; McKeever, B. M.; Ponticello, G. S..; Springer, J. P.; Randall, W. C.; Baldwin, J. J.; Habecker, C. N. Positions of His-64 and a bound water in human carbonic anhydrase II upon binding three structurally related inhibitors. *Protein Sci.* **1994**, *3*, 118−125.

(38) Boriack-Sjodin, P. A.; Zeitlin, S.; Chen, H.-H.; Crenshaw, L.; Gross, S.; Dantanarayana, A.; Delgado, P.; May, J. A.; Dean, T.; Christianson, D. W. Structural analysis of inhibitor binding to human carbonic anhydrase II. *Protein Sci.* **1998**, *7*, 2483−2489.

JM0105833